Another Generalization of Abelian Equivalence: Binomial Complexity of Infinite Words

M. Rigo[1] and P. Salimov[1,2⋆]

[1] Dept of Math., University of Liège, Grande traverse 12 (B37), B-4000 Liège, Belgium, M.Rigo@ulg.ac.be
[2] Sobolev Institute of Math., 4 Acad. Koptyug avenue, 630090 Novosibirsk, Russia.

Abstract. The binomial coefficient of two words u and v is the number of times v occurs as a subsequence of u. Based on this classical notion, we introduce the m-binomial equivalence of two words refining the abelian equivalence. The m-binomial complexity of an infinite word x maps an integer n to the number of m-binomial equivalence classes of factors of length n occurring in x. We study the first properties of m-binomial equivalence. We compute the m-binomial complexity of the Sturmian words and of the Thue–Morse word. We also mention the possible avoidance of 2-binomial squares.

1 Introduction

In the literature, many measures of complexity of infinite words have been introduced. One of the most studied is the factor complexity p_x counting the number of distinct blocks of n consecutive letters occurring in an infinite word $x \in A^{\mathbb{N}}$. In particular, Morse–Hedlund theorem gives a characterization of ultimately periodic words in terms of bounded factor complexity. Sturmian words have a null topological entropy and are characterized by the relation $p_x(n) = n + 1$ for all $n \geqslant 0$. Abelian complexity counts the number of distinct Parikh vectors for blocks of n consecutive letters occurring in an infinite word, i.e., factors of length n are counted up to abelian equivalence. Already in 1961, Erdős opened the way to a new research direction by raising the question of avoiding abelian squares in arbitrarily long words [6]. Related to Van der Waerden theorem, we can also mention the arithmetic complexity [1] mapping $n \geqslant 0$ to the number of distinct subwords $x_i x_{i+p} \cdots x_{i+(n-1)p}$ built from n letters arranged in arithmetic progressions in the infinite word x, $i \geqslant 0$, $p \geqslant 1$. In the same direction, one can also consider maximal pattern complexity [7].

As a generalization of abelian complexity, the k-abelian complexity was recently introduced through a hierarchy of equivalence relations, the coarsest being abelian equivalence and refining up to equality. We recall these notions.

⋆ The second author is supported by the Russian President's grant no. MK-4075.2012.1 and Russian Foundation for Basic Research grants no. 12-01-00089 and no. 11-01-00997 and by a University of Liège post-doctoral grant.

Let $k \in \mathbb{N} \cup \{+\infty\}$ and A be a finite alphabet. As usual, $|u|$ denotes the length of u and $|u|_x$ denotes the number of occurrences of the word x as a factor of the word u. Karhumäki *et al.* [8] introduce the notion of *k-abelian equivalence* of finite words as follows. Let u, v be two words over A. We write $u \sim_{\mathrm{ab},k} v$ if and only if $|u|_x = |v|_x$ for all words x of length $|x| \leqslant k$. In particular, $u \sim_{\mathrm{ab},1} v$ means that u and v are *abelian equivalent*, i.e., u is obtained by permuting the letters in v.

The aim of this paper is to introduce and study the first properties of a different family of equivalence relations over A^*, called k-binomial equivalence, where the coarsest relation coincide with the abelian equivalence.

Let $u = u_0 \cdots u_{n-1}$ be a word of length n over A. Let $\ell \leqslant n$. Let $t : \mathbb{N} \to \mathbb{N}$ be an increasing map such that $t(\ell - 1) < n$. Then the word $u_{t(0)} \cdots u_{t(\ell-1)}$ is a *subword* of length ℓ of u. Note that what we call subword is also called scattered subword in the literature. The notion of *binomial coefficient* of two finite words u and v is well-known, $\binom{u}{v}$ is defined as the number of times v occurs as a subword of u. In other words, the binomial coefficient of u and v is the number of times v appears as a subsequence of u. Properties of these coefficients are presented in the chapter of Lothaire's book written by Sakarovitch and Simon [12, Section 6.3]. Let $a, b \in A$, $u, v \in A^*$ and p, q be integers. We set $\delta_{a,b} = 1$ if $a = b$, and $\delta_{a,b} = 0$ otherwise. We just recall that

$$\binom{a^p}{a^q} = \binom{p}{q}, \ \binom{u}{\varepsilon} = 1, \ |u| < |v| \Rightarrow \binom{u}{v} = 0, \ \binom{ua}{vb} = \binom{u}{vb} + \delta_{a,b}\binom{u}{v}$$

and the last three relations completely determine the binomial coefficient $\binom{u}{v}$ for all $u, v \in A^*$.

Remark 1. Note that we have to make a distinction between subwords and factors. A factor is a particular subword made of consecutive letters. Factors of u are denoted either by $u_i \cdots u_j$ or $u[i, j]$, $0 \leqslant i \leqslant j < |u|$.

Definition 1. *Let $m \in \mathbb{N} \cup \{+\infty\}$ and u, v be two words over A. We say that u and v are* m-binomially equivalent *if*

$$\binom{u}{x} = \binom{v}{x}, \ \forall x \in A^{\leqslant m}.$$

Since the main relation studied in this paper is the m-binomial equivalence, we simply write in that case: $u \sim_m v$.

Since $\binom{u}{a} = |u|_a$ for all $a \in A$, it is clear that two words u and v are abelian equivalent if and only if $u \sim_1 v$. As for abelian equivalence, we have a family of refined relations: for all $u, v \in A^*$, $m \geqslant 0$, $u \sim_{m+1} v \Rightarrow u \sim_m v$.

Example 1. For instance, the four words *ababbba*, *abbabab*, *baabbab* and *babaabb* are 2-binomially equivalent. For any w amongst these words, we have the following coefficients

$$\binom{w}{a} = 3, \ \binom{w}{b} = 4, \ \binom{w}{aa} = 3, \ \binom{w}{ab} = 7, \ \binom{w}{ba} = 5, \ \binom{w}{bb} = 6.$$

But one can check that they are not 3-binomially equivalent, as an example,

$$\binom{ababbba}{aab} = 3 \text{ but } \binom{abbabab}{aab} = 4$$

indeed, for this last binomial coefficient, aab appears as subwords $w_0w_3w_4$, $w_0w_3w_6$, $w_0w_5w_6$ and $w_3w_5w_6$. Considering again the first two words, we find $|ababbba|_{ab} = 2$ and $|abbabab|_{ab} = 3$, showing that these two words are not 2-abelian equivalent. Conversely, the words $abbaba$ and $ababba$ are 2-abelian equivalent but are not 2-binomially equivalent:

$$\binom{abbaba}{ab} = 4 \text{ but } \binom{ababba}{ab} = 5.$$

This paper is organized as follows. In the next section, we present some straightforward properties of binomial coefficients and m-binomial equivalence. In Section 3, we give upper bounds on the number of m-binomial equivalence classes partitioning A^n. Section 3 ends with the introduction of the m-binomial complexity $\mathbf{b}_x^{(m)}$ of an infinite word x. In Section 4, we prove that if x is a Sturmian word then, for any $m \geqslant 2$, $\mathbf{b}_x^{(m)}(n) = n + 1$ for all $n \geqslant 0$. In Section 5 we consider the Thue–Morse word t and show that, for all $m \geqslant 1$, there exists a constant C_m such that $\mathbf{b}_t^{(m)}(n) \leqslant C_m$ for all $n \geqslant 0$. For instance, binomial coefficients of t were considered in [3]. Due to space limitations, we only give details for the cases $m = 2, 3$. In the last section, we evoke the problem of avoiding 2-binomial squares.

2 First Properties

We denote by $\mathbf{B}^{(m)}(v)$ the equivalence class of words m-binomially equivalent to v. Binomial coefficients have a nice behavior with respect to the concatenation of words.

Proposition 1. *Let p, s and $e = e_0e_1 \cdots e_{n-1}$ be finite words. We have*

$$\binom{ps}{e} = \sum_{i=0}^{n} \binom{p}{e_0e_1 \cdots e_{i-1}} \binom{s}{e_ie_{i+1} \cdots e_{n-1}}.$$

We can also mention some other basic facts on m-binomial equivalence.

Lemma 1. *Let u, u', v, v' be finite words and $m \geqslant 1$.*

- *If $u \sim_m v$, then $u \sim_\ell v$ for all $\ell \leqslant m$.*
- *If $u \sim_m v$ and $u' \sim_m v'$, then $uu' \sim_m vv'$.*

Proof. Simply note for the second point that, for all $x = x_0 \cdots x_{\ell-1}$ of length $\ell \leqslant m$, $\binom{uu'}{x}$ is equal to

$$\sum_{i=0}^{\ell} \binom{u}{x[0, i-1]} \binom{u'}{x[i, \ell-1]} = \sum_{i=0}^{\ell} \binom{v}{x[0, i-1]} \binom{v'}{x[i, \ell-1]} = \binom{vv'}{x}.$$

Remark 2. Thanks to the above lemma, we can endow the quotient set $A^*/\sim_m$ with a monoid structure using an operation $\circ : A^*/\sim_m \times A^*/\sim_m \to A^*/\sim_m$ defined by $\mathbf{B}^{(m)}(p) \circ \mathbf{B}^{(m)}(q) = \mathbf{B}^{(m)}(r)$ if the concatenation $\mathbf{B}^{(m)}(p).\mathbf{B}^{(m)}(q)$ is a subset of $\mathbf{B}^{(m)}(r)$. In particular, one can take $r = pq$. If a word v is factorized as $v = pus$, then the m-equivalence class $\mathbf{B}^{(m)}(v)$ is completely determined by p, s and $\mathbf{B}^{(m)}(u)$.

3 On the Number of k-Binomial Equivalence Classes

For 2- and 3-abelian equivalence, the number of equivalence classes for words of length n over a binary alphabet are respectively $n^2 - n + 2$ and $\Theta(n^4)$. In general, for k-abelian equivalence, the number of equivalence classes for words of length n over a ℓ-letter alphabet is $\Theta(n^{(\ell-1)\ell^{k-1}})$ [8]. We consider similar results for m-binomial equivalence (proofs are given in the appendix).

Lemma 2. *Let $u \in A^*$, $a \in A$ and $\ell \geqslant 0$. We have*

$$\binom{u}{a^\ell} = \binom{|u|_a}{\ell} \quad \text{and} \quad \sum_{|v|=\ell} \binom{u}{v} = \binom{|u|}{\ell}.$$

Lemma 3. *Let A be a binary alphabet, we have*

$$\#\left(A^n/\sim_2\right) = \sum_{j=0}^{n}((n-j)j+1) = \frac{n^3+5n+6}{6}.$$

Proposition 2. *Let $m \geqslant 2$. Let A be a binary alphabet, we have*

$$\#\left(A^n/\sim_m\right) \in \mathcal{O}(n^{2((m-1)2^m+1)}).$$

We denote by $\mathrm{Fac}_x(n)$ the set of factors of length n occurring in x.

Definition 2. *Let $m \geqslant 1$. The* m-binomial complexity *of an infinite word x counts the number of m-binomial equivalence classes of factors of length n occurring in x,*

$$\mathbf{b}_x^{(m)} : \mathbb{N} \to \mathbb{N},\ n \mapsto \#(\mathrm{Fac}_x(n)/\sim_m).$$

Note that $\mathbf{b}_x^{(1)}$ corresponds to the usual abelian complexity denoted by ρ_x^{ab}.

If p_x denotes the usual factor complexity, then for all $m \geqslant 1$, we have

$$\mathbf{b}_x^{(m)}(n) \leqslant \mathbf{b}_x^{(m+1)}(n) \quad \text{and} \quad \rho_x^{\mathrm{ab}}(n) \leqslant \mathbf{b}_x^{(m)}(n) \leqslant p_x(n). \tag{1}$$

4 The m-Binomial Complexity of Sturmian Words

Recall that a *Sturmian word x* is a non-periodic word of minimal (factor) complexity, that is, $p_x(n) = n + 1$ for all $n \geqslant 0$. The following characterization is also useful.

Theorem 1. *[13, Theorem 2.1.5] An infinite word $x \in \{0,1\}^\omega$ is Sturmian if and only if it is aperiodic and balanced, i.e., for all factors u, v of the same length occurring in x, we have $||u|_1 - |v|_1| \leqslant 1$.*

The aim of this section is to compute the m-binomial complexity of a Sturmian word as expressed by Theorem 2. We show that any two distinct factors of length n occurring in a Sturmian words are never m-binomially equivalent. First note that Sturmian words have a constant abelian complexity. Hence, if x is a Sturmian word, then $\mathbf{b}_x^{(1)}(n) = 2$ for all $n \geqslant 1$.

Theorem 2. *Let $m \geqslant 2$. If x is a Sturmian word, then $\mathbf{b}_x^{(m)}(n) = n + 1$ for all $n \geqslant 0$.*

Remark 3. If x is a right-infinite word such that $\mathbf{b}_x^{(1)}(n) = 2$ for all $n \geqslant 1$, then x is clearly balanced. If $\mathbf{b}_x^{(2)}(n) = n+1$, for all $n \geqslant 0$, then the factor complexity function p_x is unbounded and x is aperiodic. As a consequence of Theorem 2, an infinite word x is Sturmian if and only if, for all $n \geqslant 1$ and all $m \geqslant 2$, $\mathbf{b}_x^{(1)}(n) = 2$ and $\mathbf{b}_x^{(m)}(n) = n + 1$.

Before proceeding to the proof of Theorem 2, we first recall some well-known fact about Sturmian words. One of the two symbols occurring in a Sturmian word x over $\{0,1\}$ is always isolated, for instance, 1 is always followed by 0. In that latter case, there exists a unique $k \geqslant 1$ such that each occurrence of 1 is always followed by either 0^k1 or $0^{k+1}1$ and x is said to be of *type* 0. See for instance [14, Chapter 6]. More precisely, we have the following remarkable fact showing that the recoding of a Sturmian sequence corresponds to another Sturmian sequence. Note that $\sigma : A^\omega \to A^\omega$ is the shift operator mapping $(x_n)_{n\geqslant 0}$ to $(x_{n+1})_{n\geqslant 0}$.

Theorem 3. *Let $x \in \{0,1\}^\omega$ be a Sturmian word of type 0. There exists a unique integer $k \geqslant 1$ and a Sturmian word $y \in \{0,1\}^\omega$ such that $x = \sigma^c(\mu(y))$ for some $c \leqslant k + 1$ and where the morphism $\mu : \{0,1\}^* \to \{0,1\}^*$ is defined by $\mu(0) = 0^k1$ and $\mu(1) = 0^{k+1}1$.*

Corollary 1. *Let $x \in \{0,1\}^\omega$ be a Sturmian word of type 0. There exists a unique integer $k \geqslant 1$ such that any factor occurring in x is of the form*

$$0^r10^{k+\epsilon_0}10^{k+\epsilon_1}1\cdots 0^{k+\epsilon_{n-1}}10^s \tag{2}$$

where $r, s \leqslant k + 1$ and $\epsilon_0\epsilon_1\cdots\epsilon_{n-1} \in \{0,1\}^*$ is a factor of the Sturmian word y introduced in the above theorem.

Let $\epsilon = \epsilon_0\cdots\epsilon_{n-1}$ be a word over $\{0,1\}$. For $m \leqslant n-1$, we define

$$S(\epsilon, m) := \sum_{j=0}^{m}(n-j)\epsilon_j \quad \text{and} \quad S(\epsilon) := S(\epsilon, n-1). \tag{3}$$

Remark 4. Let $v = 0^r 10^{k+\epsilon_0} 10^{k+\epsilon_1} 1 \cdots 0^{k+\epsilon_{n-1}} 10^s$ of the form (2), we have

$$\binom{v}{01} = r(n+1) + \sum_{j=0}^{n-1} (k+\epsilon_j)(n-j) = r(n+1) + S(\epsilon_0 \cdots \epsilon_{n-1}) + k\frac{n(n+1)}{2}.$$

We need a technical lemma on the factors of a Sturmian word.

Lemma 4. *Let $n \geqslant 1$. If u and v are two distinct factors of length n occurring in a Sturmian word over $\{0,1\}$, then $S(u) \not\equiv S(v) \pmod{n+1}$.*

Proof. Consider two distinct factors u, v of length n occurring in a Sturmian word y. For $m < n$, we define $\Delta(m) := |u_0 u_1 \cdots u_m|_1 - |v_0 v_1 \cdots v_m|_1$. Due to Theorem 3, we have $|\Delta(m)| \leqslant 1$. Note that, if there exists i such that $\Delta(i) = 1$ then, for all $j > i$, we have $\Delta(j) \geqslant 0$. Otherwise, we would have $|v[i+1, j]|_1 - |u[i+1, j]|_1 > 1$ contradicting the fact that y is balanced. Similarly, for all $j < i$, we also have $\Delta(j) \geqslant 0$.

Since u and v are distinct, replacing u with v if needed, we may assume that there exists a minimal $i \in \{0, \ldots, n-1\}$ such that $\Delta(i) = 1$. From the above discussion and the minimality of i, $\Delta(j) = 0$ for $j < i$ and $\Delta(j) \in \{0, 1\}$ for $j > i$.

From (3), for any $j < n$, we have

$$\begin{aligned}
&\Delta(j+1) > \Delta(j) \Rightarrow S(u, j+1) - S(v, j+1) = S(u, j) - S(v, j) + (n-j)\\
&\Delta(j+1) = \Delta(j) \Rightarrow S(u, j+1) - S(v, j+1) = S(u, j) - S(v, j)\\
&\Delta(j+1) < \Delta(j) \Rightarrow S(u, j+1) - S(v, j+1) = S(u, j) - S(v, j) - (n-j).
\end{aligned}$$

In view of these observations, the knowledge of $\Delta(0), \Delta(1), \ldots$ permits to compute $(S(u, j) - S(v, j))_{0 \leqslant j < n}$ and we deduce that $0 < S(u) - S(v) < n+1$ concluding the proof.

Proof (Proof of Theorem 2). Let x be a Sturmian word of type 0 and $m \geqslant 2$. From (1), we have, for all $\ell \geqslant 0$,

$$\mathbf{b}_x^{(2)}(\ell) \leqslant \mathbf{b}_x^{(m)}(\ell) \leqslant p_x(\ell) = \ell + 1.$$

We just need to show that any two distinct factors of length ℓ in x are not 2-binomially equivalent, i.e., $\ell + 1 \leqslant \mathbf{b}_x^{(2)}(\ell)$.

Proceed by contradiction. Assume that x contains two distinct factors u and v that are 2-binomially equivalent. In particular, $\binom{u}{00} = \binom{v}{00}$ and $\binom{u}{11} = \binom{v}{11}$. Hence we get $|u| = |v|$ and $|u|_1 = |v|_1 = n$. From Corollary 1, there exist $k \geqslant 1$ and a Sturmian word y such that

$$u = 0^r 10^{k+\epsilon_0} 10^{k+\epsilon_1} 1 \cdots 0^{k+\epsilon_{n-1}} 10^s, \quad v = 0^{r'} 10^{k+\epsilon'_0} 10^{k+\epsilon'_1} 1 \cdots 0^{k+\epsilon'_{n-1}} 10^{s'}$$

where $\epsilon = \epsilon_0 \epsilon_1 \cdots \epsilon_{n-1}$ and $\epsilon' = \epsilon'_0 \epsilon'_1 \cdots \epsilon'_{n-1}$ are both factors of y.

Since $u \sim_2 v$, it follows $\binom{u}{01} = \binom{v}{01}$. From Remark 4, we get

$$r(n+1) + S(\epsilon) + k\frac{n(n+1)}{2} = r'(n+1) + S(\epsilon') + k\frac{n(n+1)}{2}.$$

Otherwise stated, we get $S(\epsilon) - S(\epsilon') = (r' - r)(n+1)$ contradicting the previous lemma.

5 The Case of the Thue–Morse Word

The *Thue–Morse* word $t = 01101001100101101001011001101001\cdots$ is the infinite word $\lim_{n\to\infty} \varphi^n(a)$ where $\varphi : 0 \mapsto 01,\ 1 \mapsto 10$. The factor complexity of the Thue–Morse word is well-known [2, 5]: $p_t(0) = 1$, $p_t(1) = 2$, $p_t(2) = 4$ and

$$p_t(n) = \begin{cases} 4n - 2\cdot 2^m - 4 \text{ if } 2\cdot 2^m < n \leqslant 3\cdot 2^m \\ 2n + 4\cdot 2^m - 2 \text{ if } 3\cdot 2^m < n \leqslant 4\cdot 2^m \end{cases}$$

and the abelian complexity of t is obvious.

Lemma 5. *We have* $\mathbf{b}_t^{(1)}(2n) = 3$ *and* $\mathbf{b}_t^{(1)}(2n+1) = 2$ *for all* $n \geqslant 1$.

The main result of this section is the following one. It is quite in contrast with the Sturmian case because here, the Thue–Morse word exhibits a bounded m-binomial complexity.

Theorem 4. *Let* $m \geqslant 2$. *There exists* $C_m > 0$ *such that the* m*-binomial complexity of the Thue–Morse word satisfies* $\mathbf{b}_t^{(m)}(n) \leqslant C_m$ *for all* $n \geqslant 0$.

For the sake of presentation, we first show that the 2-binomial complexity of the Thue–Morse word is bounded by a constant.

Theorem 5. *There exists* $C_2 > 0$ *such that the* 2*-binomial complexity of the Thue–Morse word satisfies* $\mathbf{b}_t^{(2)}(n) \leqslant C_2$ *for all* $n \geqslant 0$.

Proof. Any factor v of t admits a factorization of the kind $p\varphi(u)s$ with $p, s \in \{0, 1, \varepsilon\}$ and where u is a factor of t. Using Remark 2, it is therefore enough to prove that, for all n,

$$\#\{\mathbf{B}^{(2)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi(u)\} \leqslant 9. \tag{4}$$

Recall from the proof of Lemma 3 that the 2-binomial equivalence class of a word v of length $2n$ over a binary alphabet $\{0, 1\}$ is completely determined by its length, $|v|_0$ and $\binom{v}{01}$, i.e.,

$$\begin{aligned} &\#\{\mathbf{B}^{(2)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi(u)\} \\ &= \#\{(\binom{v}{0}, \binom{v}{1}, \binom{v}{00}, \binom{v}{01}, \binom{v}{10}, \binom{v}{11}) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi(u)\} \\ &= \#\{(|v|_0, \binom{v}{01}) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi(u)\}. \end{aligned}$$

Fix $n \geqslant 1$. Consider an arbitrary factor $u = u_0 \cdots u_{n-1} \in \mathrm{Fac}_t(n)$ and the corresponding factor $v = \varphi(u) = v_0 \cdots v_{2n-1}$ of t of length $2n$. From Lemma 5, $|v|_0$ takes at most three values (depending on n).

Let us compute the possible values taken by the coefficient $\binom{v}{01}$. Consider an occurrence of 01 as a subword of v, i.e., a pair (i, j), $i < j \leqslant n-1$, such that $v_i v_j = 01$. There are two possible cases:

– If $i = 2m$ and $j = 2m+1$, for some $m \geqslant 0$, then $u_m = 0$ because $v_{2m}v_{2m+1} = \varphi(u_m)$. There are $|u|_0$ such occurrences.
– Otherwise, we have $i \in \{2m, 2m+1\}$, $j \in \{2m', 2m'+1\}$ with $m' > m$. For all m (resp. m'), exactly one letter of the factor $v_{2m}v_{2m+1} = \varphi(u_m)$ (resp. $v_{2m'}v_{2m'+1} = \varphi(u'_m)$) is 0 and the other one is 1. Hence, for any $i \in \{0, \ldots, n-2\}$, j can take a value of the $n-1-i$ values in $\{i+1, \ldots, n-1\}$.

Summarizing these two cases, we have

$$\binom{v}{01} = |u|_0 + \sum_{i=0}^{n-2}(n-1-i) = |u|_0 + \frac{n(n-1)}{2}.$$

From Lemma 5, $|u|_0$ takes at most three values (depending on n) and therefore the same holds for $\binom{v}{01}$. Hence, the conclusion follows.

We now extend the proof of Theorem 5. The first part is to generalize (4).

Lemma 6. *Let $m, k \geqslant 1$. Assume that there exists D such that, for all n,*

$$\#\{\mathbf{B}^{(m)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi^k(u)\} \leqslant D.$$

Then the m-binomial complexity of the Thue–Morse word $\mathbf{b}_t^{(m)}$ is bounded by a constant.

Proof. Let $\ell \geqslant 1$. Let f be a factor of t of length ℓ. This factor is of the form[3] pvs where p (resp. s) is a proper suffix (resp. prefix) of some $\varphi^k(a)$ (resp. $\varphi^k(b)$) where a, b are letters and $v = \varphi^k(u)$ for some factor u of t of length n. In particular, we have $|p|, |q| \leqslant 2^k - 1$. Note that ℓ is of the form $n \cdot 2^k + r$ with $0 \leqslant r \leqslant 2(2^k - 1)$. Hence, for a given f of length ℓ, the corresponding integer n can take at most 2 values which are $\lfloor \ell/2^k \rfloor - 1$ and $\lfloor \ell/2^k \rfloor$. From the assumption, we get

$$\#\{\mathbf{B}^{(m)}(v) \mid \exists u \in \mathrm{Fac}_t(\lfloor \ell/2^k \rfloor - 1) \cup \mathrm{Fac}_t(\lfloor \ell/2^k \rfloor) : v = \varphi^k(u)\} \leqslant 2D.$$

Finally, using Remark 2, we have $\mathbf{B}^{(m)}(f) = \mathbf{B}^{(m)}(p) \circ \mathbf{B}^{(m)}(v) \circ \mathbf{B}^{(m)}(s)$. Since p and s have bounded length, $\mathbf{B}^{(m)}(p)$ and $\mathbf{B}^{(m)}(s)$ take a bounded number of values. Moreover, $\mathbf{B}^{(m)}(v)$ takes at most $2D$ values, hence $\mathbf{b}_t^{(m)}$ is bounded by constant.

From now on, intervals $[r, s]$ (resp. $[r, s)$) will be considered as intervals of integers, i.e., one should understand $[r, s] \cap \mathbb{Z}$ (resp. $[r, s) \cap \mathbb{Z}$).

Aside from the idea of dealing with words of a convenient form, the second key idea of the proof of Theorem 5 is to split the set of occurrences of the subword 01 into two disjoint subsets facilitating the counting. We shall now generalize this idea for m-binomial complexity but some terminology is required. Let v be a word. A subset $T = \{t_1 < t_2 < \ldots < t_n\} \subseteq [0, |v|)$ defines a subword denoted by $v_T = v_{t_1}v_{t_2}\cdots v_{t_n}$.

[3] This is the idea of "de-substitution" where t is factorized into consecutive factors of length 2^k.

Definition 3. *If $\alpha_1, \ldots, \alpha_m$ are non-empty and pairwise disjoint subsets of a set X such that $\cup_i \alpha_i = X$, then $\alpha = \{\alpha_1, \ldots, \alpha_m\}$ is a* partition *of X. Any partition α of a set X is a* refinement *of a partition β of X if every element of α is a subset of some element of β. In that case, α is said to be* finer *than β (equivalently β is* coarser *than α) and we write $\alpha \preceq \beta$. Since $\preceq$ is a partial order, we define a* chain *as a subset of partitions $\beta^{(1)}, \beta^{(2)}, \ldots$ of X satisfying*

$$\beta^{(1)} \preceq \beta^{(2)} \preceq \cdots .$$

*A k-*partition *$\alpha = \{\alpha_1, \ldots, \alpha_m\}$ of the set $[0, mk)$ is a partition into subsets $\alpha_i = [(i-1)k, ik)$ of size k. In particular, a 2^i-partition is a refinement of a 2^j-partition of $[0, 2^k)$, $i < j \leqslant k$.*

Definition 4. *Let X be a set and $T = \{t_1 < t_2 < \ldots < t_n\}$ be a subset of X. A partition $\alpha = \{\alpha_1, \ldots, \alpha_m\}$ of X induces a partition $\alpha_T = \{\gamma_1, \ldots, \gamma_r\}$ of $[1, n]$ defined by*

$$i, j \in \gamma_t \Leftrightarrow \exists s : t_i, t_j \in \alpha_s.$$

Note that for two partitions α, β of X, if $\alpha \preceq \beta$, then $\alpha_T \preceq \beta_T$.

Example 2. Take $X = [0, 7]$ and $T = \{0, 2, 3, 5\}$. Consider the following two partitions of X: $\alpha = \{\{0, 1\}, \{2, 3, 4\}, \{5, 6, 7\}\}$ and $\beta = \{\{0, 1, 2\}, \{3, 4, 5\}, \{6, 7\}\}$. We get $\alpha_T = \{\{1\}, \{2, 3\}, \{4\}\}$ and $\beta_T = \{\{1, 2\}, \{3, 4\}\}$.

Definition 5. *Let $T = \{t_1 < t_2 < \ldots < t_n\}$ and $U = \{u_1 < u_2 < \ldots < u_n\}$ be subsets of X. These subsets are* equidistributed *with respect to a partition α of X if $\alpha_T = \alpha_U$. These subsets are* equidistributed *with respect to a chain $\mathfrak{C}$ of partitions of X if $\alpha_T = \alpha_U$ for all $\alpha \in \mathfrak{C}$. We also say that the subsets are* $\mathfrak{C}$-equidistributed.

Example 3. Consider the chain $\mathfrak{C}$ consisting of the 4-partition $\beta = \{[0, 3], [4, 7]\}$ and the 2-partition $\alpha = \{[0, 1], [2, 3], [4, 5], [6, 7]\}$ of the set $[0, 7]$. The subsets $T = \{0, 5\}$, $U = \{1, 2\}$ and $V = \{3, 4\}$ are equidistributed with respect to the 2-partition ($\alpha_T = \alpha_U = \alpha_V = \{\{1\}, \{2\}\}$), but U is not $\mathfrak{C}$-equidistributed to T (resp. V) because $\beta_T = \beta_V = \{\{1\}, \{2\}\}$ and $\beta_U = \{\{1, 2\}\}$.

Example 4. In the last part of the proof of Theorem 5, we have considered the two possible cases for an occurrence of the subword 01 in v. If $T = \{i, j\}$ is a subset of $[0, |v|)$ and α is the 2-partition of $[0, |v|)$, then these cases correspond exactly to the two possible values $\alpha_T = \{1, 2\}$ or $\alpha_T = \{\{1\}, \{2\}\}$.

Let $\mathfrak{C}$ be a chain $\beta^{(1)} \preceq \beta^{(2)} \preceq \cdots$ of partitions of X and $T = \{t_1, \ldots, t_n\}$ be a subset of X. We use nested brackets to represent the induced chain $\beta_T^{(1)} \preceq \beta_T^{(2)} \preceq \cdots$ of partitions of $[1, n]$. The outer (resp. inner) brackets represent the coarsest (resp. finest) partition of $[1, n]$. As an example $[[t_1 t_2]][[t_3][t_4]]$ represents the partition $\{\{1, 2\}, \{3\}, \{4\}\}$ and the coarser partition $\{\{1, 2\}, \{3, 4\}\}$. To get used to these new definitions, we consider another particular statement. (A precise and formal definition of the bracket notation is given in the appendix.)

Remark 5. Two subsets T and U of size n of X are equidistributed with respect to a chain $\mathfrak{C}$ of partitions of X if and only if they give rise to the same notation of nested brackets. We call it the *type* of T with respect to $\mathfrak{C}$.

Example 5 (continuing Example 3). Consider the subsets $R = \{0, 1, 4, 7\}$ and $S = \{2, 3, 4, 6\}$ of $[0, 7]$. We have $\alpha_R = \alpha_S = \{\{1, 2\}, \{3\}, \{4\}\}$ and $\beta_R = \beta_S = \{\{1, 2\}, \{3, 4\}\}$. Hence R and S are $\mathfrak{C}$-equidistributed and give both rise to the notation $[[t_1t_2]][[t_3][t_4]]$.

We prove the case of the 3-binomial complexity. The proof of the general case needs more elaborated notions and is treated in a separated appendix.

Theorem 6. *There exists $C_3 > 0$ such that the 3-binomial complexity of the Thue–Morse word satisfies $\mathbf{b}_t^{(3)}(n) \leqslant C_3$ for all $n \geqslant 0$.*

Proof. In view of Lemma 6, it is enough to show that there exists a constant D such that, for all n, we have $\#\{\mathbf{B}^{(3)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi^2(u)\} \leqslant D$.

Let $n \geqslant 1$. Let $v = \varphi^2(u)$ with $u \in \mathrm{Fac}_t(n)$. In particular, $|v| = 4n$. Consider the chain $\mathfrak{C}$ consisting of the 2-partition and the 4-partition of $[0, 4n)$. Any subset $T = \{t_1 < t_2 < t_3\}$ of $[0, 4n)$ is $\mathfrak{C}$-equidistributed to a subset of one the following types:

- $[t_1][t_2][t_3]$, i.e., the union of the types $[[t_1]][[t_2]][[t_3]]$, $[[t_1][t_2]][[t_3]]$ and $[[t_1]][[t_2][t_3]]$: the 3 elements of T belong to pairwise distinct subsets of the 2-partition of $[0, 4n)$
- $[[t_1t_2][t_3]]$ or $[[t_1][t_2t_3]]$: two elements belong to the same subset of the 2-partition of $[0, 4n)$ and the 3 elements of T belong to the same subset of the 4-partition of $[0, 4n)$.
- $[[t_1t_2]][[t_3]]$ or $[[t_1]][[t_2t_3]]$: two elements belong to the same subset of the 2-partition and to the same subset of the 4-partition of $[0, 4n)$.

Let $e = e_0e_1e_2$ be a word of length 3. We will count the number of occurrences of the subword $e = v_{t_1}v_{t_2}v_{t_3}$ in v depending on the type of $T = \{t_1, t_2, t_3\}$ with respect to $\mathfrak{C}$.

Assume that the type of T is $[t_1][t_2][t_3]$. Each subset S of the 2-partition of $[0, 4n)$ corresponds to a factor $v_S = 01$ or $v_S = 10$ and v contains $2n$ such factors. Hence the number of subwords e occurring in v for this type takes, for a given n, a unique value which is $\binom{2n}{3}$.

Now assume that the type of T is $[[t_1t_2][t_3]]$ (similar arguments apply to $[[t_1][t_2t_3]]$). Each subset S of the 4-partition of $[0, 4n)$ corresponds to a factor v_S which is either $\varphi^2(0) = 0110$ or $\varphi^2(1) = 1001$. Then the number of subwords e occurring in v of this type is

$$\underbrace{\binom{01}{e_0e_1}}_{0 \text{ or } 1}\underbrace{\binom{10}{e_2}}_{1}|u|_0 + \underbrace{\binom{10}{e_0e_1}}_{0 \text{ or } 1}\underbrace{\binom{01}{e_2}}_{1}|u|_1 \in \{0, |u|_0, |u|_1\}.$$

Recall that, for a given $n = |u|$, the pair $(|u|_0, |u|_1)$ can take at most three values (see Lemma 5). The number of subwords e occurring in v of this type takes, for a given n, takes at most 4 values[4].

Now assume that the type of T is $[[t_1t_2]][[t_3]]$ (similar arguments apply to $[[t_1]][[t_2t_3]]$). Each subset S of the 4-partition of $[0, 4n)$ is a union of two sets S', S'' of the 2-partition of $[0, 4n)$ and we have either $v_{S'} = 01, v_{S''} = 10$ or $v_{S'} = 10, v_{S''} = 01$. They are n subsets of size 4 in the 4-partition of $[0, 4n)$ and we have to pick 2 of them. Hence, the number of subwords e occurring in v for this type is

$$(\underbrace{\binom{01}{e_0e_1} + \binom{10}{e_0e_1}}_{0 \text{ or } 1})(\underbrace{\binom{01}{e_2} + \binom{10}{e_2}}_{2})\binom{n}{2}$$

and this quantity, for a given n, takes at most 2 values.

We have proved that, for all $|e| = 3$ and $v = \varphi^2(u)$ with $u \in \mathrm{Fac}_t(n)$, $\binom{v}{e}$ takes at most $1 + 2 \cdot 4 + 2 \cdot 2 = 13$ values (these values depend on n, but the *number* of values is bounded without any dependence to n). Note that $\mathbf{B}^{(3)}(v)$ is determined from $\mathbf{B}^{(2)}(v)$ and by the values of $\binom{v}{e}$ for the words e of length 3. To conclude the proof, note that $\#\{\mathbf{B}^{(2)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi^2(u)\}$ is bounded by $\#\{\mathbf{B}^{(2)}(v) \mid \exists z \in \mathrm{Fac}_t(2n) : v = \varphi(z)\} \leqslant 9$ using (4). Consequently, we have shown that $\#\{\mathbf{B}^{(3)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi^2(u)\} \leqslant 9 \cdot 13^8$ for all $n \geqslant 1$.

Remark 6. By computer experiments, $\mathbf{b}_t^{(2)}(n)$ is equal to 9 if $n \equiv 0 \pmod 4$ and to 8 otherwise, for $10 \leqslant n \leqslant 1000$. Moreover, $\mathbf{b}_t^{(3)}(n)$ is equal to 21 if $n \equiv 0 \pmod 8$ and to 20 otherwise, for $8 \leqslant n \leqslant 500$.

6 A Glimpse at Avoidance

It is obvious that, over a 2-letter alphabet, any word of length $\geqslant 4$ contains a square. On the other hand, there exist square-free infinite ternary words [12]. In the same way, over a 3-letter alphabet, any word of length $\geqslant 8$ contains an abelian square, i.e., a word uu' where $u \sim_1 u'$. But, over a 4-letter alphabet, abelian squares are avoidable, see for instance [10]. So a first natural question in that direction is to determine, whether or not, over a 3-letter alphabet 2-binomial squares can be avoided in arbitrarily long words. Naturally, a 2-*binomial square* is a word of the form uu' where $u \sim_2 u'$. Note that, for abelian equivalence, the longest ternary word which is 2-abelian square-free has length 537 [9].

As an example, $u = 121321231213123132123121312$ is a word of length 27 without 2-binomial squares but this word cannot be extended without getting a 2-binomial square. Indeed, $u1$ (resp. $u3$) ends with a square of length 8 (resp. 26)

Consider the 13-uniform morphism of Leech [11] which is well-known to be square-free, $g : a \mapsto abcbacbcabcba, b \mapsto bcacbacabcacb, c \mapsto cabacbabcabac$. In

[4] A close inspection shows that if $|u| = 2n$, then $|u|_0, |u|_1 \in \{n-1, n, n+1\}$, if $|u| = 2n+1$, then $|u|_0, |u|_1 \in \{n, n+1\}$.

the submitted version of this paper, we conjectured that the infinite square-free word $g^\omega(1)$ avoids 2-binomial squares. For instance, we can prove that

$$u \sim_2 v \Leftrightarrow g(u) \sim_2 g(v).$$

Nevertheless, M. Bennett has recently shown that the factor of length 508 occurring in position 845 is a 2-binomial square [4].

Acknowledgments

The idea of this binomial equivalence came after the meeting "Representing streams" organized at the Lorentz center in December 2012 where Jean-Eric Pin presented a talk, *Noncommutative extension of Mahlers theorem on interpolation series*, involving binomial coefficients on words. Jean-Eric Pin and the first author proposed independently to introduce this new relation.

References

1. S.V. Avgustinovich, D.G. Fon-Der-Flaass, A.E. Frid, Arithmetical complexity of infinite words, in *Words, Languages & Combinatorics III*, M. Ito and T. Imaoka (Eds.), World Scientific Publishing (2003) 51–62.
2. S. Brlek, Enumeration of factors in the Thue-Morse word, *Discrete Appl. Math.* **24** (1989), 83–96.
3. J. Berstel, M. Crochemore, J.-E. Pin, Thue–Morse sequence and p-adic topology for the free monoid, *Disc. Math.* **76** (1989), 89–94.
4. J. Currie, personal communication, 3th June 2013.
5. A. de Luca, S. Varricchio, On the factors of the Thue-Morse word on three symbols, *Inform. Process. Lett.* **27** (1988), 281–285.
6. P. Erdős, Some unsolved problems, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6** (1961), 221-254.
7. T. Kamae, L. Zamboni, Sequence entropy and the maximal pattern complexity of infinite words, *Ergodic Theory Dynam. Systems* **22** (2002), 1191–1199.
8. J. Karhumäki, A. Saarela, L. Q. Zamboni, On a generalization of Abelian equivalence and complexity of infinite words, `arXiv:1301.5104`.
9. M. Huova, J. Karhumäki, Observations and problems on k-abelian avoidability, In Combinatorial and Algorithmic Aspects of Sequence Processing (Dagstuhl Seminar 11081), (2011) 2215-2219.
10. V. Keränen, Abelian squares are avoidable on 4 letters, *Lecture Notes in Comput. Sci.* **623** (1992), 41-52.
11. J. Leech, A problem on strings of beads, *Math. Gazette* **41**, 277–278 (1957).
12. M. Lothaire, *Combinatorics on Words*, Cambridge Mathematical Library, Cambridge University Press, (1997).
13. M. Lothaire, *Algebraic Combinatorics on Words*, Encyclopedia of Mathematics and its Applications **90**, Cambridge University Press (2002).
14. N. Pytheas Fogg, *Substitutions in dynamics, arithmetics and combinatorics*, V. Berthé, S. Ferenczi, C. Mauduit and A. Siegel (Eds.), Lecture Notes in Mathematics **1794**, Springer-Verlag, Berlin, 2002.

7 Proofs on the Number of Equivalence Classes

Proof (Proof of Lemma 3). The set A^n is split into $n+1$ equivalence classes for the abelian equivalence. Let $j \in \{0, \ldots, n\}$. Consider a representative u of such an abelian equivalence class characterized by $|u|_a = j$ and $|u|_b = n - j$. The extremal values taken by $\binom{u}{ab}$ are

$$\binom{b^{n-j}a^j}{ab} = 0 \text{ and } \binom{a^j b^{n-j}}{ab} = j(n-j).$$

Now we show that, for all $k \in \{0, \ldots, j(n-j)\}$, there exists a word w abelian equivalent to u and such that $\binom{w}{ab} = k$. One has simply to consider the $j(n-j)+1$ words $b^{n-j}a^j$, $b^{n-j-1}aba^{j-1}$, $b^{n-j-1}a^2ba^{j-2}$, $\ldots$, $b^{n-j-1}a^jb$, $b^{n-j-2}aba^{j-1}b$, $b^{n-j-2}a^2ba^{j-2}b$, $\ldots$, $b^{n-j-2}a^jb^2$, $\ldots$, a^jb^{n-j}. To conclude the proof, since $\binom{w}{aa}$ and $\binom{w}{bb}$ are determined by the abelian class of w and using Lemma 2

$$\binom{w}{ba} = \binom{|w|}{2} - \binom{w}{aa} - \binom{w}{bb} - \binom{w}{ab}$$

then, the coefficient $\binom{w}{ba}$ for a word w abelian equivalent to u is deduced from $\binom{w}{ab}$. Hence the abelian equivalence class containing $b^{n-j}a^j$ is split into $(n-j)j+1$ classes for the 2-binomial equivalence.

Proof (Proof of Proposition 2). Let u be a word of length n. Let $\ell \leqslant m$. The number of subwords of length ℓ occurring in u is $\binom{n}{\ell}$. There are exactly 2^ℓ words of length ℓ enumerated lexicographically: $v_{\ell,1}, \ldots, v_{\ell,2^\ell}$. Consider the vector $\Psi_\ell(u)$ of size 2^ℓ given by

$$\Psi_\ell(u) := \left(\binom{u}{v_{\ell,1}} \cdots \binom{u}{v_{\ell,2^\ell}} \right).$$

If u and u' are two words of length n such that $\Psi_\ell(u) \neq \Psi_\ell(u')$, then $u \not\sim_\ell u'$ and thus $u \not\sim_m u'$. An upper bound on the number of values taken by $\Psi_\ell(u) \in \mathbb{N}^{2^\ell}$ is given by the number of ways to partition the integer $\binom{n}{\ell}$ as a sum of 2^ℓ non-negative integers, that is $(\binom{n}{\ell}+1)^{2^\ell-1}$. Hence, we get

$$\#\left(A^n/\sim_m\right) \leqslant \prod_{\ell=1}^{m} (\binom{n}{\ell}+1)^{2^\ell-1}.$$

The upper bound is obtained by replacing $(\binom{n}{\ell}+1)^{2^\ell-1}$ with $(n^\ell)^{2^\ell}$.

8 Binomial Complexity of the Thue–Morse Word

From the proof of Theorem 6, when dealing with the general case of m-binomial complexity of the Thue–Morse word t, it seems clear that we have to study the possible types of occurrences of a subword of length m with respect to a chain of $2-, 4-, 8-, \ldots, 2^m-$partitions of $[0, 2^m n)$. To describe these types in full generality, we introduce some notation and more elaborated definitions. To help the reader, we have considered along the text several examples illustrating the different new notions.

8.1 $\mathfrak{C}$-Equidistributed n-Subsets

For the set X, we usually have in mind the set $[0, |v|)$ of positions inside a word v.

Definition 6. *Let X be a set and $n \leqslant \#X$. A* n-set *is a set of size n. Let $\mathfrak{C}$ be a chain of partitions of X. Being $\mathfrak{C}$-equidistributed is obviously an equivalence relation $\equiv_{\mathfrak{C}}$ over the set X_n of the n-subsets of X. The corresponding quotient set is denoted by $X_n/\equiv_{\mathfrak{C}}$.*

Example 6. Consider the set $X = [0, 7]$ and the chain $\mathfrak{D}$ consisting of the 2-partition and the 4-partition of X. The set

$$\mathcal{E} = \{\{0 + i_1, 2 + i_2, 4 + i_3, 6 + i_4\} \mid i_1, \ldots, i_4 \in \{0, 1\}\}$$

is an equivalence class belonging to the quotient set $X_4/\equiv_{\mathfrak{D}}$. In other words, a 4-subset of X belongs to $\mathcal{E}$ if and only if its four elements belong to pairwise distinct subsets of the 2-partition of X. In the same way, the set

$$\mathcal{F} = \{\{0 + i_1, 1 + i_1, i_2\} \mid i_1 \in \{0, 2\}, i_2 \in \{4, 5, 6, 7\}\}$$

is an equivalence class belonging to $X_3/\equiv_{\mathfrak{D}}$.

Definition 7. *Let k, n be integers such that $n \geqslant k \geqslant 1$. Let $S = \{s_1, \ldots, s_k\}$ be a subset of $[1, n]$ and $\mathcal{E} \in X_n/\equiv_{\mathfrak{C}}$ be an equivalence class of $\mathfrak{C}$-equidistributed n-subsets of X. The* restriction *of $\mathcal{E}$ to S, denoted by $\mathcal{E}|_S$, is the set of $\mathfrak{C}$-equidistributed k-subsets of X defined by*

$$U = \{u_1 < \cdots < u_k\} \in \mathcal{E}|_S \Leftrightarrow \exists T = \{t_1 < \cdots < t_n\} \in \mathcal{E}, \forall i \in [1, k] : u_i = t_{s_i}.$$

We say that k is the size *of $\mathcal{E}|_S$. It is the cardinal of any subset belonging to $\mathcal{E}|_S$. If $\mathcal{R}$ is a restriction of the kind $\mathcal{E}|_S$, then* $\mathrm{dom}(\mathcal{R})$ *denotes the underlying set S.*

Example 7 (Example 6 cont.). Let $S = \{1, 3\}$. We have

$$\mathcal{E}|_S = \{\{0, 4\}, \{0, 5\}, \{1, 4\}, \{1, 5\}\}.$$

Note that $\mathcal{E}|_S$ is a strict subset of the equivalence class

$$\{\{i, j\} \mid i \in [0, 3], j \in [4, 7]\} \in X_2/\equiv_{\mathfrak{D}} .$$

With the same set S, we get $\mathcal{F}|_S = \{\{i, j\} \mid i \in \{0, 2\}, j \in [4, 7]\}$.

Definition 8. *Let $\mathfrak{C} = \{\alpha^{(1)} \preceq \alpha^{(2)} \preceq \cdots \preceq \alpha^{(r)}\}$ be a chain of partitions of X and $\alpha \in \mathfrak{C}$. Let $\mathcal{E}$ be an equivalence class in $X_n/\equiv_{\mathfrak{C}}$. Recall that, for all n-subsets $U, T \in \mathcal{E}$, we have $\alpha_U = \alpha_T$. The* partition of $[1, n]$ induced *by $\mathcal{E}$ is denoted by $\alpha_{\mathcal{E}}$ and is equal to α_T for any $T \in \mathcal{E}$.*

Consider the coarsest partition $\alpha^{(r)} \in \mathfrak{C}$. Then $\alpha^{(r)}_{\mathcal{E}}$ is a partition $\{S_1, \ldots, S_t\}$ of $[1, n]$. The restrictions $\mathcal{E}|_{S_1}, \ldots, \mathcal{E}|_{S_t}$ are called the components *of $\mathcal{E}$.*

If $\mathcal{E}|_{S_1}, \ldots, \mathcal{E}|_{S_t}$ are the components of $\mathcal{E}$, then any subset T belonging to $\mathcal{E}$ admits a unique decomposition as

$$T = U_1 \cup \cdots \cup U_t, \quad \text{with } U_j \in \mathcal{E}|_{S_j}, \forall j.$$

Remark 7. As seen in Example 7, a restriction and thus a component of an equivalence class $\mathcal{E} \in X_n/\equiv_{\mathfrak{C}}$ is generally a strict subset of an equivalence class in $X_k/\equiv_{\mathfrak{C}}$ for some k.

Example 8 (Example 6 cont.). Consider again the chain $\mathfrak{D}$ consisting of the 2-partition $\alpha^{(1)} = \{[0,1],[2,3],[4,5],[6,7]\}$ and the 4-partition $\alpha^{(2)} = \{[0,3],[4,7]\}$ of $[0,7]$. Take the 4-subset $T = \{0,2,4,6\}$ which is a class representative of $\mathcal{E}$. We have $\alpha_T^{(1)} = \alpha_{\mathcal{E}}^{(1)} = \{\{1\},\{2\},\{3\},\{4\}\}$ and $\alpha_T^{(2)} = \alpha_{\mathcal{E}}^{(2)} = \{S_1, S_2\}$ where $S_1 = \{1,2\}$ and $S_2 = \{3,4\}$. The components of $\mathcal{E}$ are

$$\mathcal{E}|_{S_1} = \{\{0,2\},\{0,3\},\{1,2\},\{1,3\}\} \text{ and } \mathcal{E}|_{S_1} = \{\{4,6\},\{4,7\},\{5,6\},\{5,7\}\}.$$

Consider the 3-subset $U = \{0,1,4\}$ which is a class representative of $\mathcal{F}$. We have $\alpha_U^{(1)} = \alpha_{\mathcal{F}}^{(1)} = \{\{1,2\},\{3\}\}$ and $\alpha_U^{(2)} = \alpha_{\mathcal{F}}^{(2)} = \{\{1,2\},\{3\}\}$. The components of $\mathcal{F}$ are $\{\{0,1\},\{2,3\}\}$ and $\{\{4\},\{5\},\{6\},\{7\}\}$.

8.2 The Dot-Bracket Notation

We now introduce a convenient way to denote a class of $\mathfrak{C}$-equidistributed n-subsets of X. Roughly speaking, this notation represents how the n elements of any n-subset of the equivalence class are distributed inside the partitions of $\mathfrak{C}$.

Definition 9. *Let $\mathfrak{C} = \{\alpha^{(1)} \preceq \alpha^{(2)} \preceq \cdots \preceq \alpha^{(r)}\}$ be a chain of partitions of X. Let $\mathcal{E}$ be an equivalence class in $X_n/\equiv_{\mathfrak{C}}$. We introduce the* dot-bracket notation *for $\mathcal{E}$ as the word $db(\mathcal{E})$ over the alphabet $\{[,],\cdot\}$, containing exactly n dots and defined inductively as follows:*

- *Step 0: define n words $g(0,1) = \cdots = g(0,n) = \cdot$ (corresponding to the singletons $\{1\},\ldots,\{n\}$). We set $\alpha_{\mathcal{E}}^{(0)} = \{\{1\},\ldots,\{n\}\}$.*
- *From step t to step $t+1$: we have s_t words $g(t,1),\ldots,g(t,s_t)$ corresponding to the partition of $[1,n]$ given by $\alpha_{\mathcal{E}}^{(t)} = \{S_1^{(t)},\ldots,S_{s_t}^{(t)}\}$. Assume that $\alpha_{\mathcal{E}}^{(t+1)} = \{S_1^{(t+1)},\ldots,S_{s_{t+1}}^{(t+1)}\}$. We define s_{t+1} new words $g(t+1,1),\ldots,g(t+1,s_{t+1})$ such that*

$$g(t+1,j) = [g(t,i)g(t,i+1)\cdots g(t,i+k)] \text{ if } S_j^{(t+1)} = S_i^{(t)} \cup \cdots \cup S_{i+k}^{(t)}.$$

The size *of a dot-bracket notation $M \in \{[,],\cdot\}^*$ is the number of occurrences of "$\cdot$" in the word M.*

Example 9 (Example 6 cont.). We get $db(\mathcal{E}) = [[\cdot][\cdot]]\,[[\cdot][\cdot]]$ and $db(\mathcal{F}) = [[\cdot\cdot]]\,[[\cdot]]$.

Remark 8. Words over $\{[,],\cdot\}$ representing an equivalence class in $X_n/\equiv_{\mathfrak{C}}$ are of a special form. They correspond to Motzkin paths of a special kind, i.e., lattice paths of $\mathbb{N}^2$ running from $(0,0)$ to $(t,0)$, for some t, that never pass below the x-axis and whose permitted steps are the up diagonal step $(1,1)$ corresponding to "[", the down diagonal step $(1,1)$ corresponding to "]" and the horizontal step $(1,0)$ corresponding to "$\cdot$", satisfying the following properties:

- each subpath connecting $(i,0)$ and $(k,0)$ and not passing through $(j,0)$, $i<j<k$, has height $\#\mathfrak{C}$,
- there are exactly n horizontal steps and they all occur at height $\#\mathfrak{C}$,
- the word does not contain the factor $[\,]$, i.e., there is no peak made of an up diagonal step followed by a down diagonal step.

We extend the dob-bracket notation, defined above for an equivalence class, to any subset of an equivalence class (and in particular to its components). This is meaningful because the considered n-subsets of X are still $\mathfrak{C}$-equidistributed. Hence, this notation represents how the elements of any such n-subset are distributed with respect to the partitions of $\mathfrak{C}$.

Definition 10. *Let $\mathcal{E}|_{S_1},\dots,\mathcal{E}|_{S_t}$ be the components of $\mathcal{E}\in X_n/\equiv_{\mathfrak{C}}$. Let $j\in\{1,\dots,t\}$. The elements of $\mathcal{E}|_{S_j}$ are ℓ_j-subsets of $[1,n]$ and $\mathcal{E}|_{S_j}$ is included in some equivalence class $\mathcal{F}_j\in X_{\ell_j}/\equiv_{\mathfrak{C}}$. Note that $\ell_1+\cdots+\ell_t=n$. (Observe that $\mathrm{dom}(\mathcal{E}|_{S_j})=[1+\ell_1+\cdots+\ell_{j-1},\ell_1+\cdots+\ell_j]$ for all $j\leqslant t$.) We define the* dot-bracket notation *for the component $\mathcal{E}|_{S_j}$ as $db(\mathcal{F}_j)$. Note that the dot-bracket notation of the components can be deduced from $db(\mathcal{E})$: there are exactly the factors of the kind $[^{\#\mathfrak{C}}u]^{\#\mathfrak{C}}$, where u does not contain the factor $[^{\#\mathfrak{C}}$, occurring in $db(\mathcal{E})$. By abuse of language, these factors are called the* components *of $db(\mathcal{E})$ and we can speak of the* size *of a component.*

Example 10 (Example 6 cont.). The two components of $db(\mathcal{E})$ are both equal to $[[\cdot][\cdot]]$ and the two components of $db(\mathcal{F})$ are $[[\cdot\cdot]]$ and $[[\cdot]]$. For instance, the notation $[[\cdot\cdot]]$ means that the two elements of any subset of the component belong to the same 2-partition and to the same 4-partition. Note that this notation for a component lacks the information about the underlying domain. The component denoted by $[[\cdot\cdot]]$ corresponds to the distribution of the two subsets $\{0,1\}$ and $\{2,3\}$ inside $\mathfrak{D}$. But with the equivalence class $\mathcal{G}=\{\{i_1,4+i_2,5+i_2\}\mid i_1\in\{0,1,2,3\},i_2\in\{0,2\}\}$ denoted by $db(\mathcal{G})=[[\cdot]]\,[[\cdot\cdot]]$, the component denoted again by $[[\cdot\cdot]]$ corresponds to the distribution of the two subsets $\{4,5\}$ and $\{6,7\}$ inside $\mathfrak{D}$.

Note that the same dot-bracket notation may occur for equivalence classes related to two different sets and their corresponding chains of partitions. We will often make use of this fact in all what follows. For instance, when considering a component and then the equivalence class having the same dot-bracket notation.

Definition 11. *Let X be a set, $\mathfrak{C}$ be a chain of partitions of X and $\mathcal{E}$ be an equivalence class in $X_m/\equiv_{\mathfrak{C}}$. Let Y be a set, $\mathfrak{D}$ be a chain of partitions of Y such that $\#\mathfrak{C}=\#\mathfrak{D}$.*

We denote by $\mathrm{cl}(Y, \mathfrak{D}, db(\mathcal{E}))$ *the equivalence class* $\mathcal{F}$ *in* $Y_m/\equiv_{\mathfrak{D}}$ *such that* $db(\mathcal{E}) = db(\mathcal{F})$*. If no such class exists or if* $\#\mathfrak{C} \neq \#\mathfrak{D}$*, we set* $\mathrm{cl}(Y, \mathfrak{D}, db(\mathcal{E})) = \emptyset$*.*

Example 11. Let $\mathfrak{C}$ (resp. $\mathfrak{C}'$) be the chain consisting of 2- and 4-partitions of $[0, 4)$ (resp. $[0, 8)$). Let $\mathfrak{D}$ be the chain consisting of 3- and 9-partitions of $[0, 9)$. For instance, we have

$$\mathrm{cl}([0,4), \mathfrak{C}, [[\cdot][\cdot]]) = \{\{0+i_1, 2+i_2\} \mid i_1, i_2 \in \{0,1\}\},$$

$$\mathrm{cl}([0,8), \mathfrak{C}', [[\cdot][\cdot]]) = \{\{0+i_1, 2+i_2\}, \{4+i_1, 6+i_2\} \mid i_1, i_2 \in \{0,1\}\},$$

$$\mathrm{cl}([0,9), \mathfrak{D}, [[\cdot][\cdot]]) = \{\{0+i_1, 3+i_2, 6+i_3\} \mid i_1, i_2, i_3 \in \{0,1,2\}\}.$$

8.3 Extending the Chain of Partitions

Let T and U be two n-subsets that are $\mathfrak{D}$-equidistributed for some chain $\mathfrak{D}$ of partitions of X. If we add an extra partition α, there is no reason that $\alpha_T = \alpha_U$ holds true. Therefore, the equivalence class in $X_n/\equiv_{\mathfrak{D}}$ containing U and T will be split into several classes with respect to $\equiv_{\mathfrak{D}\cup\{\alpha\}}$. A straightforward consequence of the previous definitions is the following proposition. See Example 12 below for an illustration of this result.

Proposition 3. *Let* $\mathfrak{C}, \mathfrak{D}$ *be two chains of partitions of* X *such that* $\mathfrak{C} = \mathfrak{D} \cup \{\alpha\}$ *and* α *being coarser than any partition in* $\mathfrak{D}$*. Let* $\mathcal{E} \in X_n/\equiv_{\mathfrak{D}}$ *be an equivalence class having a dot-bracket notation factorized with its components* $db(\mathcal{E}) = C_1 \cdots C_s$*. The class* $\mathcal{E}$ *is a union of* $r \geqslant 1$ *classes* $\mathcal{F}_1, \ldots, \mathcal{F}_r$ *belonging to* $X_n/\equiv_{\mathfrak{C}}$ *such that, for all* $i \leqslant r$*:*

(D1) either, $db(\mathcal{F}_i) = [C_1][C_2]\cdots[C_s]$*, or*
(D2) $db(\mathcal{F}_i)$ *contains less components than* $db(\mathcal{E})$*.*

The classes $\mathcal{F}_1, \ldots, \mathcal{F}_r$ are referred to the *sons* of $\mathcal{E}$ (whenever $\mathfrak{C}, \mathfrak{D}$ are clear from the context) and we write $sons(\mathcal{E}) = \{\mathcal{F}_1, \ldots, \mathcal{F}_r\}$. We use the same terminology for the corresponding dot-bracket notation. The sons of $db(\mathcal{E})$ are $db(\mathcal{F}_1), \ldots, db(\mathcal{F}_r)$. Assume, for some $t \geqslant 1$, that the word $M = N_1 \cdots N_t$ over $\{[,], \cdot\}$ is a dot-bracket notation factorized with its components where $N_i = [u_i]$ for all i. Then $u_1 \cdots u_t$ is a dot-bracket notation P where M is a son of P.

To highlight components, we will make use of larger brackets.

Example 12. Consider the set $X = [0, 15]$ and the chain $\mathfrak{D}$ consisting of the 2- and 4-partitions of X. Let α be the 8-partition of X and $\mathfrak{C} = \mathfrak{D} \cup \{\alpha\}$. Let $\mathcal{E}$ be the class in $X_3/\equiv_{\mathfrak{D}}$ such that $db(\mathcal{E}) = \big[[\cdot]\big]\,\big[[\cdot\cdot]\big]$. For instance, the sets $\{0, 6, 7\}$ and $\{0, 8, 9\}$ belong to $\mathcal{E}$. But these two sets belong to two different equivalence classes with respect to $\equiv_{\mathfrak{C}}$, $\mathcal{F}_1$ and $\mathcal{F}_2$ respectively where

$$db(\mathcal{F}_1) = \big[[[\cdot]][[\cdot\cdot]]\big],\ db(\mathcal{F}_2) = \big[[[\cdot]]\big]\,\big[[[\cdot\cdot]]\big] \text{ and } \mathcal{E} = \mathcal{F}_1 \cup \mathcal{F}_2.$$

The reason is that the elements of any subset in $\mathcal{E}$ belong to exactly two of the four subsets making the 4-partition of $X = [0, 15]$. There is some freedom left

when choosing the two subsets of the 4-partition: they can be subsets of the same subset of the 8-partition of X (leading to $\mathcal{F}_1$) or they can be subsets of the two distinct subsets of the 8-partition of X (leading to $\mathcal{F}_2$). To illustrate the previous proposition, note that $db(\mathcal{F}_1)$ contains less components than $db(\mathcal{E})$ and $db(\mathcal{F}_2)$ satisfies (D1). As another example, take $\mathcal{E}'$ be the class in $X_6/\equiv_{\mathfrak{D}}$ such that $db(\mathcal{E}') = \big[[\cdot][\cdot]\big]\ \big[[\cdot]\big]\ \big[[\cdot][\cdot]\big]\ \big[[\cdot]\big]$. This class is also a single class of $X_6/\equiv_{\mathfrak{C}}$ with dot-bracket notation $\big[[[\cdot][\cdot]][[\cdot]]\big]\ \big[[[\cdot][\cdot]][[\cdot]]\big]$. Here we do not have the same freedom as in the previous case: the elements of any subset in $\mathcal{E}'$ belong to exactly the four subsets making the 4-partition of $X = [0, 15]$.

Definition 12. *Let v and $e = e_0e_1\cdots e_{m-1}$ be two finite words. Let $\mathfrak{C}$ be a chain of partitions of $X = [0, |v|)$. Let $\mathcal{E}$ be an equivalence class in $X_m/\equiv_{\mathfrak{C}}$. We denote by*

$$\binom{v}{e}_{\mathcal{E}}$$

the number of m-subsets T of $[0, |v|)$ such that $T \in \mathcal{E}$ and $v_T = e$. Note that if the size of $\mathcal{E}$ and the length of e differ, then for any $T \in \mathcal{E}$, $v_T \neq e$ and $\binom{v}{e}_{\mathcal{E}} = 0$.

More generally, if $N \in \{[,], \cdot\}^*$ is a dot-bracket notation (in particular, the dot-bracket notation of a component) and if the chain $\mathfrak{C}$ is understood from the context, then we set

$$\binom{v}{e}_{N} := \binom{v}{e}_{\mathrm{cl}([0,|v|),\mathfrak{C},N)}.$$

Thanks to the definitions, the computation of $\binom{v}{e}$ can be split as follows:

$$\binom{v}{e} = \sum_{\mathcal{E}\in X_m/\equiv_{\mathfrak{C}}} \binom{v}{e}_{\mathcal{E}}.$$

Example 13 (Example 6 cont.). Consider the set $X = [0, 7]$ and the equivalence class $\mathcal{E} = \{\{0 + i_1, 2 + i_2, 4 + i_3, 6 + i_4\} \mid i_1, \ldots, i_4 \in \{0, 1\}\}$. Let v be the prefix 01101001 of the Thue–Morse word and $e = 0101$. We have

$$\binom{v}{e}_{\mathcal{E}} = \binom{v}{e}_{[[\cdot][\cdot]]\,[[\cdot][\cdot]]} = 1.$$

Indeed, there is a single $T = \{0, 2, 5, 7\} \in \mathcal{E}$ such that $v_T = e$.

8.4 About the Thue–Morse Word

Since the Thue–Morse Word is generated by the 2-uniform morphism φ, we will consider particular chains of partitions. We denote by $\mathfrak{C}^{(i)}$ the chain consisting of the 2^j-partitions (of a convenient set) for all $1 \leqslant j \leqslant i$. With now ambiguity, we use a single notation for the chain $\mathfrak{C}^{(i)}$ of partitions of any set $[0, 2^in)$.

Now we state the key technical observations specific to the Thue–Morse word generated by φ. Note that in [12, p. 140], the following statement is given as an exercise: for all $n \geqslant 0$ and all word f of length less than n,

$$\binom{\varphi^n(0)}{f} = \binom{\varphi^n(1)}{f}.$$

We extend this kind of result to binomial coefficient with respect to a class.

From now on, X is assumed to be $[0, 2^i t)$ with large enough t.

Proposition 4. *Let $i, m \geqslant 1$. Let N be a dot-bracket notation of size m. If N has a single component then, for all words e of length m and all $k > 0$, we have*

$$\binom{\varphi^{i+k}(0)}{e}_N = \binom{\varphi^{i+k}(1)}{e}_N \tag{5}$$

i.e.,

$$\binom{\varphi^{i+k}(0)}{e}_{\mathrm{cl}([0,2^{i+k}),\mathfrak{C}^{(i)},N)} = \binom{\varphi^{i+k}(1)}{e}_{\mathrm{cl}([0,2^{i+k}),\mathfrak{C}^{(i)},N)}$$

Proof. Let $\mathcal{E}$ be the equivalence class $\mathrm{cl}([0, 2^{i+k}), \mathfrak{C}^{(i)}, N)$. Since $\mathcal{E}$ has a single component, each set T in $\mathcal{E}$ is a subset of an element of the 2^i-partition of $[0, 2^{i+k})$. Observe that for any element S of the 2^i-partition of $[0, 2^{i+k})$ and for $v = \varphi^{i+k}(0)$ or $v = \varphi^{i+k}(1)$, we always have $v_S \in \{\varphi^i(0), \varphi^i(1)\}$. Moreover, $\varphi^{i+k}(0)$ and $\varphi^{i+k}(1)$ are both factorized as a product of 2^{k-1} words $\varphi^i(0)$ and 2^{k-1} words $\varphi^i(1)$ and the conclusion follows. Note that only this last observation is specific to the morphism φ.

In following statements, we will make use of the abbreviated form given for instance in (5) because the underlying set and the chain can easily be deduced from the context. In the following lemma, the equivalence class can have more than one component.

Lemma 7. *Let $i, m \geqslant 1$. Let $\mathcal{E} \in X_m/\equiv_{\mathfrak{C}^{(i)}}$. If $\mathcal{E}$ has r components $\mathcal{C}_1, \ldots, \mathcal{C}_r$ such that, for all all $j \in \{1, \ldots, r\}$ and all words f of length equal to the size of $\mathcal{C}_j$*

$$\binom{\varphi^i(0)}{f}_{db(\mathcal{C}_j)} = \binom{\varphi^i(1)}{f}_{db(\mathcal{C}_j)},$$

then, for all words e of length m and all words u, we have

$$\binom{\varphi^i(u)}{e}_{db(\mathcal{E})} = \binom{|u|}{r} \prod_{j=1}^{r} \binom{\varphi^i(0)}{e_{\mathrm{dom}(\mathcal{C}_j)}}_{\mathrm{cl}([0,2^i),\mathfrak{C}^{(i)},db(\mathcal{C}_j))}.$$

Proof. We take $X = [0, 2^i|u|)$. We choose r subsets $S_1, \ldots, S_r$ amongst the $|u|$ sets constituting the 2^i-partition of X. For any of these subsets S, $v_S \in \{\varphi^i(0), \varphi^i(1)\}$ where $v = \varphi^i(u)$. Moreover, note that the chain $\mathfrak{C}^{(i)}$ of partitions of X induces, by intersection, a chain of partitions of each S_j. We denote again this chain $\mathfrak{C}^{(i)}$.

Each possible choice of $S_1, \ldots, S_r$ corresponds exactly to the m-subsets $T \in \mathcal{E}$ such that $T \subseteq S_1 \cup \cdots \cup S_r$. For such a choice, we consider an occurrence of e as $e = e_{\mathrm{dom}(\mathcal{C}_1)} \cdots e_{\mathrm{dom}(\mathcal{C}_r)}$ where $e_{\mathrm{dom}(\mathcal{C}_j)}$ occurs in v_{S_j} with respect to $\mathcal{C}_j$. For all j, there are

$$\binom{v_{S_j}}{e_{\mathrm{dom}(\mathcal{C}_j)}}_{\mathrm{cl}([0,2^i),\mathfrak{C}^{(i)},db(\mathcal{C}_j))}$$

such occurrences. Thanks to the assumption, we do not have to distinguish between the case $v_{S_j} = \varphi^i(0)$ and $v_{S_j} = \varphi^i(1)$ and the conclusion follows.

Note that in Proposition 4, (5) holds for exponents of φ larger than i, but in the assumption of Lemma 7, the relation (5) should hold for an exponent equal to i. This is why we consider the next two results.

Lemma 8. *Let $i \geqslant 2$, $m \geqslant 1$. Let $\mathcal{E}' \in X_m/\equiv_{\mathfrak{C}^{(i-1)}}$ be an equivalence class having a single component. Let $\mathcal{E} \in X_m/\equiv_{\mathfrak{C}^{(i)}}$ be such that $db(\mathcal{E}) = [db(\mathcal{E}')]$. Then $\mathcal{E}' = \mathcal{E}$ and, for all words e of length m and for all $k \geqslant i$, we have*

$$\binom{\varphi^k(0)}{e}_{\mathrm{cl}([0,2^k),\mathfrak{C}^{(i)},db(\mathcal{E}))} = \binom{\varphi^k(1)}{e}_{\mathrm{cl}([0,2^k),\mathfrak{C}^{(i)},db(\mathcal{E}))}.$$

Proof. Note that from the assumption on $\mathcal{E}$, each set T in $\mathcal{E}$ is a subset of an element of the 2^{i-1}-partition of $[0, 2^k)$. One can then follows the proof of Proposition 4.

Since we are dealing with the two chains $\mathfrak{C}^{(i-1)}$ and $\mathfrak{C}^{(i)}$, we can consider a specific instance of Proposition 3.

Lemma 9. *Let $i \geqslant 2$, $m \geqslant 1$. Let $\mathcal{E} \in X_m/\equiv_{\mathfrak{C}^{(i-1)}}$ having a dot-bracket notation factorized with its components $db(\mathcal{E}) = C_1 \cdots C_s$. The class $\mathcal{E}$ is a union of $r \geqslant 1$ classes $\mathcal{F}_1, \ldots, \mathcal{F}_r$ belonging to $X_m/\equiv_{\mathfrak{C}^{(i)}}$.*

Assume that $\mathcal{F}_j$ satisfies (D1), i.e., $db(\mathcal{F}_j) = [C_1][C_2]\cdots[C_s]$. Then $\mathcal{F}_j$ satisfies the assumptions of Lemma 7: for all $r \in \{1, \ldots, s\}$, for all words f of length equal to the size of $[C_r]$, we have

$$\binom{\varphi^i(0)}{f}_{[C_r]} = \binom{\varphi^i(1)}{f}_{[C_r]}.$$

Proof. Observe that any T in $\mathrm{cl}([0, 2^i), \mathfrak{C}^{(i)}, [C_r])$ is a subset of an element of the 2^{i-1}-partition of $[0, 2^i)$. One can then follows the proof of Proposition 4.

We can now turn to the proof of the main result about the Thue–Morse word.

Theorem 7. *Let $m \geqslant 2$. There exists $C_m > 0$ such that the m-binomial complexity of the Thue–Morse word satisfies $\mathbf{b}_t^{(m)}(n) \leqslant C_m$ for all $n \geqslant 0$.*

Proof. We proceed by induction on m. The cases $m = 2, 3$ have already been considered. In particular, we may assume that, for all n,

$$\#\{\mathbf{B}^{(m-1)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi^{m-1}(u)\} \leqslant \mathbf{b}^{(m-1)}(2^{m-1}n) \leqslant C_{m-1}.$$

Hence, we already have, for all n,

$$\begin{aligned} &\#\{\mathbf{B}^{(m-1)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi^m(u)\} \\ \leqslant\ &\#\{\mathbf{B}^{(m-1)}(v) \mid \exists z \in \mathrm{Fac}_t(2n) : v = \varphi^{m-1}(z)\} \leqslant C_{m-1}. \end{aligned}$$

In view of Lemma 6, it is enough to show that there exists a constant D_m such that, for all n, we have $\#\{\mathbf{B}^{(m)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi^m(u)\} \leqslant D_m$. Since $\mathbf{B}^{(m)}$ is determined by $\mathbf{B}^{(m-1)}(v)$ and by $\binom{v}{e}$ for all words e of length m. We just have to concentrate of the number of values that can be taken by $\binom{v}{e}$ in that case.

Let $n \geqslant 1$. We have to prove that, for all $|e| = m$ and $v = \varphi^m(u)$ with $u \in \mathrm{Fac}_t(n)$, $\binom{v}{e}$ takes a number of values bounded by a constant (these values depend on n, but the *number* of values does not depend on n). In particular, $|v| = 2^m n$.

For all $i \geqslant 0$, considering the chain $\mathfrak{C}^{(i)}$ of partitions of $X = [0, 2^m n)$, we have

$$\binom{v}{e} = \sum_{\mathcal{E} \in X_m/\equiv_{\mathfrak{C}^{(i)}}} \binom{v}{e}_{\mathcal{E}}$$

where $\mathfrak{C}^{(0)}$ contains only the partition of $[0, 2^m n)$ split into singletons. Hence $X_m/\equiv_{\mathfrak{C}^{(0)}}$ contains a unique class $\mathcal{R}$ consisting of all the m-subsets of X. We set $db(\mathcal{R}) = ([\cdot])^m$ and

$$\binom{v}{e} = \binom{v}{e}_{\mathcal{R}}. \tag{6}$$

Let $i \geqslant 0$. Since an equivalence class $\mathcal{E} \in X_m/\equiv_{\mathfrak{C}^{(i)}}$ is the union of its sons which are equivalence classes in $X_m/\equiv_{\mathfrak{C}^{(i+1)}}$, we trivially have

$$\binom{v}{e}_{\mathcal{E}} = \sum_{\mathcal{F} \in sons(\mathcal{E})} \binom{v}{e}_{\mathcal{F}}. \tag{7}$$

Note that $\mathcal{R}$ does not satisfy the conditions of Lemma 7.

We will build a tree of height bounded by m and having $\mathcal{R}$ as root on level 0. On level i, the nodes are some equivalence classes in $X_m/\equiv_{\mathfrak{C}^{(i)}}$. For a given word e, values will be attached to the leaves of that tree. Thanks to (6) and (7), $\binom{v}{e}$ will be given by the sum of the leaves of the tree. The construction is given as follows:

- If a node $\mathcal{E} \in X_m/\equiv_{\mathfrak{C}^{(i)}}$ on level i has r components $\mathcal{C}_1, \ldots, \mathcal{C}_r$ and satisfies the conditions of Lemma 7, then this node is a leaf. The attached valuation of the leaf is

$$\binom{v}{e}_{\mathcal{E}} = \binom{|u^{(i)}|}{r} \prod_{j=1}^{r} \binom{\varphi^i(0)}{e_{\mathrm{dom}(\mathcal{C}_j)}}_{\mathrm{cl}([0,2^i),\mathfrak{C}^{(i)},db(\mathcal{C}_j))} \tag{8}$$

  where the word $u^{(i)}$ is such that $\varphi^i(u^{(i)}) = v = \varphi^m(u)$. In particular, $|u^{(i)}| = 2^{m-i}n$.
- If a node $\mathcal{E} \in X_m/\equiv_{\mathfrak{C}^{(i)}}$ on level i does not satisfy the conditions of Lemma 7, then it is not a leaf. We consider its sons on level $i+1$. For each son $\mathcal{F}$ of $\mathcal{E}$, we face the alternative given in Proposition 3:
  - Either, $\mathcal{F}$ satisfies (D1). Then, from Lemma 9, $\mathcal{F}$ satisfies the conditions of Lemma 7 and we are back to the first situation: $\mathcal{F}$ is a leave.

- Or, $\mathcal{F}$ satisfies (D2) and this son has less component than $\mathcal{E}$. Again, for such a node on level $i+1$, one has to check if the conditions of Lemma 7 are satisfied. Note that, since the number of components in each equivalence class is decreasing, with at most $m-1$ steps, all the remaining sons not satisfying the conditions of Lemma 7 will be reduced to single component. In that latter case, making use of Lemma 8, Lemma 7 can eventually be applied to the only remaining son.

Before proceeding to the conclusion of the proof, we consider an example for constructing this tree when $m = 3$ (we take n large enough). The tree is depicted in Figure 1. The nodes $\mathcal{R}_{1,1}$, $\mathcal{R}_{2,1}$ and $\mathcal{R}_{2,3}$ are sons satisfying (D1).

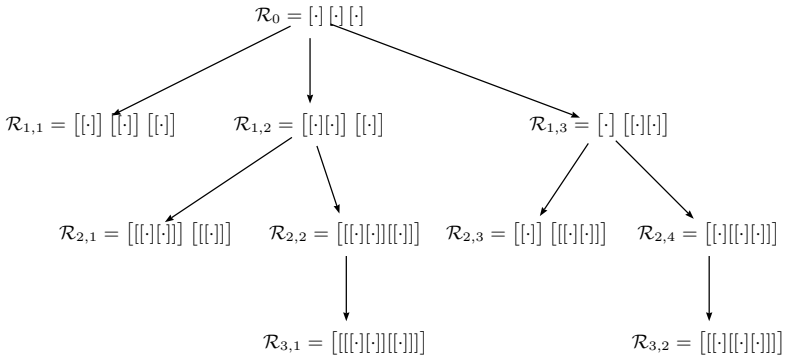

Fig. 1. The tree when $m = 3$.

These nodes are leafs, Lemma 7 can be applied. The nodes $\mathcal{R}_{1,2}$, $\mathcal{R}_{1,3}$, $\mathcal{R}_{2,2}$ and $\mathcal{R}_{2,4}$ are sons satisfying (D2). Note that on the two paths starting from $\mathcal{R}_0$ and leading respectively to $\mathcal{R}_{2,2}$ and $\mathcal{R}_{2,4}$, the number of components decreases on each level. The nodes $\mathcal{R}_{2,2}$ and $\mathcal{R}_{2,4}$ have a single component. Thanks to Lemma 8, their respective sons $\mathcal{R}_{3,1}$ and $\mathcal{R}_{3,2}$ are satisfying the assumptions of Lemma 7. Note as a general observation that the leafs correspond exactly to equivalence classes where each component has a dot-bracket notation of the kind $[[z]]$ where z is a well-parenthesed word. One may observe that the structure of this tree depends only on m.

Now to conclude with the proof, observe that, m being a constant, what may only change are the valuations (8) of the leafs.

Assume that n is fixed. When v is a word from the set $\{\varphi^m(u) \mid u \in \mathrm{Fac}_t(n)\}$ and e is any word of length m, the possible values of $\binom{v}{e}$ are obtained from the valuations of the leafs. But, in (8), $|u^{(i)}| = 2^{m-i}n$ depends only on n. Hence the valuations of the leafs given by (8) depends only on n, e. This means that for a given n, $\binom{v}{e}$ can take at most 2^m different values (the number of pairwise distinct values is bounded by a constant, but the values are depending on n).

9 Partial Results on the Leech Morphism

For palindromes, we have the following relations. The reversal of a word u is denoted by u^R.

Lemma 10. *Let u, v be words and a, b, c be letters. If $v = uu^R$, then*

$$\binom{v}{ab} = \binom{u}{ab} + \binom{u}{ba} + |u|_a|u|_b$$

and if $v = ucu^R$, then

$$\binom{v}{ab} = \binom{u}{ab} + \binom{u}{ba} + \delta_{a,b}|u|_a|uc|_b + (1-\delta_{a,b})|uc|_a|uc|_b.$$

Consequently, if two palindromes are abelian equivalent, then they are 2-binomially equivalent.

Proof. For the first formula, the first (resp. second and third) term counts the number of occurrences of the subword ab in the first half u of the word v (resp. in the second half u^R of v and finally in v with a occurring in the first half u and b in the second half u^R).

Here g refers to the Leech's morphism. Consider a variation of the map Ψ_ℓ introduced in the proof of Proposition 2. For a word $u \in \{a, b, c\}$,

$$\begin{aligned}\Psi(u) = \; & (|u|_a, |u|_b, |u|_c, \tbinom{u}{aa}, \tbinom{u}{ab}, \tbinom{u}{ac}, \\ & \tbinom{u}{ba}, \tbinom{u}{bb}, \tbinom{u}{bc}, \tbinom{u}{ca}, \tbinom{u}{cb}, \tbinom{u}{cc})^\top.\end{aligned}$$

We now introduce a matrix $M_g \in \mathbb{N}^{12\times 12}$ such that $M\Psi(u) = \Psi(g(u))$

$$M_g = \begin{pmatrix}
4 & 4 & 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 4 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 5 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
6 & 6 & 10 & 16 & 16 & 20 & 16 & 16 & 20 & 20 & 20 & 25 \\
10 & 8 & 10 & 20 & 16 & 16 & 20 & 16 & 16 & 25 & 20 & 20 \\
8 & 10 & 10 & 16 & 20 & 16 & 16 & 20 & 16 & 20 & 25 & 20 \\
10 & 8 & 10 & 20 & 20 & 25 & 16 & 16 & 20 & 16 & 16 & 20 \\
10 & 6 & 6 & 25 & 20 & 20 & 20 & 16 & 16 & 20 & 16 & 16 \\
10 & 10 & 8 & 20 & 25 & 20 & 16 & 20 & 16 & 16 & 20 & 16 \\
8 & 10 & 10 & 16 & 16 & 20 & 20 & 20 & 25 & 16 & 16 & 20 \\
10 & 10 & 8 & 20 & 16 & 16 & 25 & 20 & 20 & 20 & 16 & 16 \\
6 & 10 & 6 & 16 & 20 & 16 & 20 & 25 & 20 & 16 & 20 & 16
\end{pmatrix}$$

Note that the upper-left 3×3 corner is the usual matrix associated with g. For instance, the fifth line is obtained as follows. To get $\binom{g(u)}{ab}$ from $\Psi(u)$, we have to count $\binom{g(x)}{ab}$ for each symbol $x \in \{a, b, c\}$ but we have also to take into account the subwords ab obtained by taking a symbol a in a block $g(x)$ and a symbol b in

another block $g(y)$. Since all letters have 4 or 5 occurrences in every block $g(x)$ of length 13, this explains the values 16, 20 and 25. This matrix is invertible and therefore

$$\begin{aligned} u \sim_2 v &\Leftrightarrow \Psi(u) = \Psi(v) \Leftrightarrow M_g\Psi(u) = M_g\Psi(v) \\ &\Leftrightarrow \Psi(g(u)) = \Psi(g(v)) \Leftrightarrow g(u) \sim_2 g(v). \end{aligned}$$